High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering

Alain Lefebvre[1,*], Emilie Poisson-Caillault[1,2]

[1]Ifremer, Laboratoire Environnement et Ressources, 62200 Boulogne-sur-Mer, France
[2]LISIC, EA 4491, Université du Littoral Côte d'Opale, 62228 Calais, France

ABSTRACT: As we move towards shipboard-underway and automated systems for monitoring water quality and assessing ecological status, there is a need to evaluate how effective the existing monitoring systems are, and how we could improve them. Considering the existing limitations for processing numerous and complex data series generated from automated systems, and because of processes involved in phytoplankton blooms, this paper proposes a data-driven evaluation of an unsupervised classifier to optimize the way we track phytoplankton, including harmful algal blooms (HABs), and to identify the main associated hydrological conditions. We used *in situ* data from a portable flow-through automatic measuring system coupled with a multi-fixed-wavelength fluorometer implemented in the eastern English Channel during a bloom of *Phaeocystis globosa* (high biomass, non-toxic HAB species). This combination of technologies allowed high resolution online hydrographical and biological measurements, including spectral fluorescence as a means of quantifying phytoplankton biomass and simplifying the phytoplankton community structure inference. An unsupervised spectral clustering method was applied to this multi-parameter high-resolution time series, which allowed discrimination under near real-time of 6 to 33 contrasting water masses based on their abiotic and biotic characteristics. In addition, areas subject to extreme events such as HABs could be precisely identified, so controlling factors or their direct and indirect effects could be hierarchized. Considering the benefits and limitations of such a strategy, future applications of such methods will be important in the context of implementing the Marine Strategy Framework Directive.

KEY WORDS: Clustering · Spectral fluorescence · *Phaeocystis globosa* · English Channel · Ferry Box · MSFD · Marine Strategy Framework Directive

1. INTRODUCTION

Processes involved in phytoplankton bloom dynamics cover a wide range of spatio-temporal scales (Cloern 1996, Dickey 2003); bloom events may only last a few days, but can have strong direct and indirect effects on ecosystems. The lack of high resolution *in situ* monitoring systems for water quality assessment is critical given that the eastern English Channel (EEC) is classified as 'moderate' to 'poor'

with respect to the Water Framework Directive's (WFD) phytoplankton quality element. The French coastal zone of the EEC is also classified as a potential eutrophication problem area when implementing the Oslo and Paris Convention (OSPAR) Common Procedure (OSPAR Commission 2010). Indeed, the French part of the EEC is regularly affected by *Phaeocystis globosa* bloom events, leading to massive foam production and accumulation, with negative direct and indirect effects on benthic and pelagic

components of the ecosystem (Lancelot 1995, Rousseau et al. 2006). *P. globosa* is classified as a high biomass, nontoxic harmful algal bloom (HAB) species, and *P. globosa* blooms are often concomitant to *Pseudo-nitzschia* sp. events, when paralytic shellfish poisoning could occur (Sazhin et al. 2007). Phytoplankton dynamics are particularly sensitive to anthropogenic pressures (Boyce et al. 2010, Hernández-Fariñas et al. 2014) and this biological compartment and the resulting primary production play a fundamental role in influencing the dynamics and diversity of other biological compartments. HABs are at the core of important health and environmental issues (Smayda 1990, Cloern & Jassby 2008) and there has been a significant increase in their occurrence worldwide (Heisler et al. 2008, Glibert 2017). This responsiveness of phytoplankton to environmental change (from eutrophication to global changes) and to the ecosystem goods and services associated with it explain why phytoplankton is widely used as an indicator for the development of environmental quality assessment systems (European Parliament and Council 2000, 2008, OSPAR Commission 2009).

The Marine Strategy Framework Directive (MSFD) is the legal framework for the future management of Europe's marine environment. It requires that the EU achieve Good Environmental Status (GES) of its marine waters by 2020, due to the implementation of a monitoring programme. The Directive defines GES as 'the environmental status of marine waters where these provide ecologically diverse and dynamic oceans and seas which are clean, healthy and productive' (European Parliament and Council 2008, p. 25).

In general, observation and monitoring of highly dynamic systems like coastal waters requires intensive sampling in space and time to detect short-term events which might have a strong impact on the ecosystem (Petersen et al. 2006). Focusing on the English Channel–North Sea MSFD sub-region, we suspect that existing observations and monitoring programmes in the French part of the EEC lack the high spatio-temporal resolution coverage required to determine the state of the marine environment and changes therein. Unfortunately, this is also the case for other EU marine waters where there is a lack of standardized and accessible observational data in offshore areas (Crise et al. 2015, Teixeira et al. 2016). Most of the regular and continuous observations and monitoring programmes in the EEC only cover the coastal zone. The number of stations drastically declines or is nonexistent offshore (beyond the 20 m isobath) (French MSFD initial assessment 2012; http://

sextant.ifremer.fr/en/web/dcsmm/pamm/evaluation-initiale). Whereas the French part of the EEC is of strong ecological and socio-economic importance, only 2 areas are equipped with a high temporal resolution system—a fixed platform based on buoy technology: MAREL Carnot, since 2004 in the northern part of the EEC (50° 44′ 26″ N, 1° 34′ 4″ E) (Zongo & Schmitt 2011, Zongo et al. 2011, Rousseeuw et al. 2015), and the SMILE buoy implemented in 2015 in the Bay of Seine (49° 20′ 37″ N, 0° 18′ 27″ W).

In the EuroMarine Strategy Research report, Boyen et al. (2012) highlighted several issues. (1) Do we really understand marine ecosystems well enough to use present knowledge (mainly based on low frequency, conventional studies) in such a context? (2) What are the relevant advances in knowledge and where are the gaps? (3) What are the new technologies being required to observe and understand marine ecosystems?

Since ecological studies and overall ecological assessments need to integrate physical, chemical and biological information to provide a comprehensive view of the states and dynamics of the ecosystems, and since they need to be in line with their main spatial and temporal controlling scales (Cloern 1996, Mantua et al. 2002), we decided to implement an integrated, autonomous, multi-sensor system called Pocket Ferry Box (PFB), coupled with a multiple-fixed-wavelength spectral fluorometer (Algae Online Analyser [AOA]; bbe Moldaenke).

This study takes place within the framework of research projects devoted to the development of dynamic observation systems for assessing marine water quality based on high resolution phytoplankton analysis (www.dymaphy.eu, www.jerico-ri.eu). The main objective was to develop and/or implement a high spatial and temporal resolution measuring system, then to assess its added value considering conventional approaches (low resolution with mainly delayed laboratory analysis) in order to deepen knowledge on phytoplankton bloom dynamics. It deals more generally with the study of marine water quality at appropriate spatial and temporal scales: either we assess phytoplankton indicators, or we consider the MSFD issues for descriptors on biodiversity and eutrophication.

Our paper was built as a proof-of-concept to characterize simple phytoplankton communities using high resolution measurement devices coupled with optimized numerical methodologies for data processing. The main objectives of our paper were:

(1) Considering numerical methodological development issues, to test the performance of several unsu-

pervised classification methods in order to propose the most optimized one to comply with label-specific environmental conditions as a way to discriminate phytoplankton spectral groups and associated hydrological conditions;

(2) From an ecological viewpoint, to characterize high resolution spatio-temporal variability of phytoplankton spectral fluorescence in a temperate, well-mixed open coastal system. We also studied the relationships between spectral fluorescence and major abiotic variables in order to evaluate the usefulness of such techniques to identify environmental conditions, eco-regions, and the main factors controlling phytoplankton structure, as well as direct and indirect effects of blooms events on ecosystems, making a link with eutrophication and HAB issues. Considering the complex hydrodynamics of the studied area and the specificity of such an environment partly dominated by the Prymnesiophyceae *P. globosa* in spring, we hypothesised that there would be (i) a high spatio-temporal variability of phytoplankton spectral fluorescence and total chlorophyll *a* (chl *a*) concentration, and (ii) a strong difference in phytoplankton fluorescence signatures between UK and French waters;

(3) From a technical viewpoint, to identify advantages and drawbacks of spectral fluorometry, using a limited number of phytoplankton spectral groups, as a preliminary taxonomic approach.

The added value of such a system for continuous monitoring purposes and environmental indicator development was assessed while considering the need to obtain a thorough understanding of optimal spatio-temporal scales controlling phytoplankton blooms. Moreover, we also anticipated and discussed the implementation of the monitoring programme for the EU directives (i.e. MSFD) and regional sea conventions (i.e. OSPAR).

2. MATERIALS AND METHODS

2.1. Sampling sites and strategies

The EEC is characterized by a macrotidal regime, with a tidal range of ~3 and ~9 m during neap and spring tides, respectively, in the Dover Strait (Anonymous 1988). The latter generates fast tidal currents essentially parallel to the coast and a northeast-flowing tidal residual current from the English Channel to

the North Sea (Salomon & Breton 1993). Along the French coast in the EEC, fluvial supplies from the Bay of Seine to Cap Gris-Nez (including the Bay of Somme, which is the second most important main river at the English Channel scale, and some other minor contributors) generate a coastal water mass that drifts nearshore, separated from the open sea by a frontal area. This frontal area plays an important role in structuring exchanges (non-biological and biological particles) between coastal and offshore water masses (Brylinski & Lagadeuc 1990). Along the southern UK coast of the English Channel, the main rivers are the Avon and the Arun, and the riverine inputs are consequently lesser than those on the French side. In the EEC no seasonal pycnocline occurs; where freshwater influence is prevalent, stratification is scarce and intermittent depending on levels of freshwater discharge.

Data were collected during DYMAPHY ((Development of a DYnamic observation system for the assessment of MArine water quality, based on PHYtoplankton analysis) cruises onboard the RV 'Côtes de la Manche' (CNRS/INSU) between 20 and 21 April 2012 (Leg 1), between 27 and 29 April 2012 (Leg 2) and between 31 May and 4 June 2012 (Leg 3) (Fig. 1). During these 3 legs, we investigated water
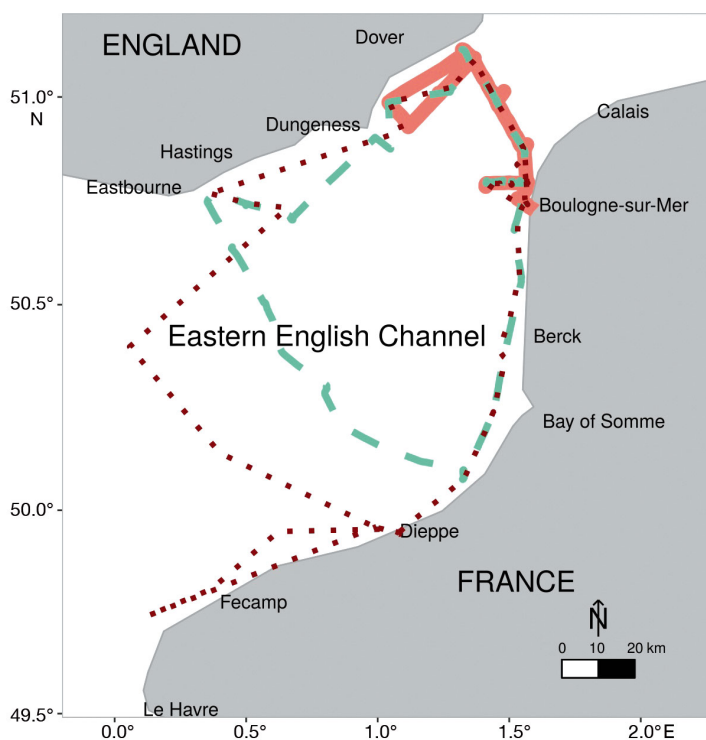

Fig. 1. Location of the 3 sampling routes (Leg 1: bold orange line; Leg 2: broken green line; Leg 3: dotted red line) in the eastern English Channel for cruises during spring 2012

masses at high resolution (<200 m) (see Section 2.2) along routes approximately 100, 200 and 300 nautical miles long, respectively. During the cruises, water samples were regularly collected from sub-surface waters as ground truth reference results using conventional analytical procedures (see 'Low resolution sample analysis' section below).

2.2. High resolution sample analysis

A detailed description of the PFB system (4H-JENA) implemented during the cruises was made by Schroeder et al. (2008) and Aiken et al. (2011). For our campaigns, the water intake was at the front of the ship's cooling system at a fixed depth (3 m); in normal ship operation seawater is constantly pumped. The non-filtered seawater flow through the PFB was set to be, on average, between 3 and 4 l min^{-1}. All parameters were recorded at a frequency of 1 min^{-1} and geographical coordinates were provided by an external GPS receiver. The RV 'Côtes de la Manche' sailed at approximately 6 knots during the cruises, so the spatial resolution of recording was about 185 m. Because of the relatively brief transit time of water from the water intake to the PFB, the observations are representative of sub-surface conditions. The system used during this project was assembled with sensors for salinity and temperature (conductivity sensor 3919; AANDERA), pH (MV 3010; Meinsberg), oxygen concentration (oxygen optode 3835; AANDERA) and coloured dissolved organic matter (CDOM; Cyclops 7 sensor; AANDERA). Unfortunately, poor quality of pH and CDOM data did not allow data processing as planned. All sensors were quality checked by the manufacturer before the cruise.

The PFB was coupled with a multiple-fixed-wavelength spectral fluorometer, the AOA (bbe Moldaenke), which continuously measures the chlorophyll fluorescence of microalgae under real time (results are expressed in eq. µg chl a l^{-1}). The measurement corresponds to wet chemical chlorophyll analysis (DIN/ISO) and is based on determination of the fluorescence spectrum and fluorescence kinetics of the algae. All measurements in the AOA are based on the fluorescence characteristics of chlorophyll and other pigments of the algae (Yentsch & Yentsch 1979, Ruser et al. 1999, Beutler et al. 2002). The algae classes can be differentiated by their pigments and therefore by the different fluorescence responses of their differing photosystem antenna systems and accessory pigments to light of various wavelengths. Each algae class has its own fluorescence signature or fingerprint, which is

a special pattern according to which they react to different excitation wavelengths (470, 525, 570, 590 and 610 nm). With this technique, the chl a content of green algae (G) (Chlorophyta), blue-green algae (BG) (*Cyanobacteria* with phycocyanin), Cryptophyceae (C) (Cryptophyta, Rhodophyta, *Cyanobacteria* with phycoerythrin) and brown algae (B) (Heterokontophyta and Dinophyta) (in AOA parlance) was analysed. To compensate for the influence of turbidity on fluorescence, transmission measurement was used as a tool for correction (370 nm).

Our AOA was calibrated by the manufacturer using cultures of *Chlorella vulgaris*, *Microcystis aeruginosa*, *Cyclotella meneghiniana* and *Cryptomonas* sp., respectively for G, BG, B and C signals. During calibration, the photosystem was excited with visible light, one wavelength after another. Between excitations, the red fluorescence emission was measured. So the response of all phytoplankton for this excitation wavelength was collected. From the calibration, the fluorescence intensity of 1 µg l^{-1} of each group was deduced for each excitation wavelength. The AOA software allocates the fluorescence emission intensity to the fingerprint of the algae groups and makes the best fit. The resulting ratio of fluorescence to chl a (F:chl) is used to convert the fluorescence data into units of chl a (eq. µg chl a l^{-1}). Deviations between reference concentrations and actual values were 0.96, 0.26, 0.13 and 2.39%, respectively for G, BG, B and C signals. The linear model based on the relationship between active chl a measured by fluorescence by AOA (Chl_AOA) and total chl a measured by spectrometry analysis (Chl_Spec) confirmed that the AOA signal could be used as a proxy for chl a even if it is better to consider the overall trends, the qualitative aspect rather than the quantitative one (Chl_AOA = 2.13 × Chl_Spec + 0.15, with adjusted R^2 = 0.55, p < 0.001, n = 23).

2.3. Low resolution sample analysis

As in Lefebvre et al. (2011), water samples were collected from sub-surface waters using a 5 l Niskin bottle. Each hydrological sampling station was referenced using an internal code (identified as DPM or Stn) followed by a number. Chl a concentrations were estimated by spectrophotometry (Shimadzu UV-1700) after filtration through Whatman 47 mm GF/F glass fiber filters and after extraction in 90% acetone (Aminot & Kérouel 2004).

Phytoplankton samples were collected from the Niskin bottle and preserved with an acid Lugol's

solution. Sub-samples of 10 ml were settled for 24 h in a counting chamber according to the Utermöhl method (Utermöhl 1958). Cell enumerations were performed by inverted microscopy within 1 mo of sample collection to prevent any significant changes in phytoplankton size and abundance. Except for *Phaeocystis globosa* enumeration, over 400 phytoplankton cells in each sample were counted with a 20× Plan Ph1 0.5NA objective, resulting in an error of 10%. For assessment of *P. globosa* counts, only the total number of cells was computed. A minimum of 50 solitary cells were enumerated from several randomly chosen fields (10 to 30) with a 40× Plan Ph2 0.75NA. The abundance of cells in a colony was determined using the relationship between colony biovolume and cell number defined by Rousseau et al. (1990) (see the Supplement at www.int-res.com/articles/suppl/m608p073_supp.xlsx for details about phytoplankton cell enumeration for each reference discrete sampled station).

2.4. Data analysis

The basic statistical analyses applied here are detailed in Legendre & Legendre (1998) and Shumway & Stoffer (2006). All statistical analyses were performed using R statistical software (R Development Core Team 2009). The main R packages used were 'uHMM', 'kmeans', 'hclust', 'mclust' and 'e1071'.

Different types of features were used as inputs for the classification system:

• Set 1, named AOA: 4 raw algae concentrations from AOA (G, BG, B and Cryptophyceae).

• Set 2, named Prop: proportions of the 4 algae groups derived from the raw algae concentrations from Set 1 and the chlorophyll (sum of the 4 spectral groups). This set is similar to the one used by the marine water manager to see the principal ruptures in the water composition and to detect algae blooms, in particular the presence of *Phaeocystis* algae.

• Set 3, named PFB: 4 raw hydrological parameters from the PFB—salinity, sea surface water temperature, oxygen concentration and oxygen saturation.

• Set 4, named AOA-PFB: concatenation of AOA and PFB (Sets 1+3).

Sensors were selected considering the range in the variables expected in the area. Regardless of the set, data outside of the sensor's range were removed (NA). Then, an algorithm of detection of outliers or aberrant data was processed. It consisted of erasing features (by NA) when the value was out of the 1−99% range or when the consecutive gap divided by both the boat speed and the median speed was greater than 2× the upper whisker. For any classifier, data features were centred and scaled by the standard deviations. Time information and longitude–latitude information were not included in the clustering process.

Three clustering methods were compared in order to optimize extraction of the maximum amount of information contained in this high-resolution time series: hierarchical clustering (HC), Hartigon *k*-means (*k*-means) and NJW-spectral clustering (SC). HC was included due to its popularity in the marine community; here, the Ward's minimum variance method (Ward 1963, Murtagh & Legendre 2014) was chosen. Labels were obtained by cutting the resulting dendogram at the desired *k* number of clusters. The *k*-means approach was preferred over the more general expectation–maximization (EM) algorithm approach both for its fast computing time and its single parameter, *k*, to tune. EM requires a choice of the covariance structures, and has no real added-value on result improvement, according to Rousseeuw et al. (2015). *k*-means is based on minimization of the within-cluster sum of squares of each cluster. It is well adapted to globular shape data, whereas EM is preferred for ellipsoid shape data with different centres. Here, the R package 'kmeans' was used (with Hartigon options if convergence was assumed, else with the McQueen algorithm). SC (Ng et al. 2001) is able to model and separate clusters with different shapes and the same center of distributions, but it applies the *k*-means algorithms not directly on the raw space but on the spectral subspace. This subspace is composed of the *k* eigenvectors extracted from the Laplacian matrix based on the Zelnik and Perona similarity (Zelnik-Manor & Perona 2004). The SC algorithm is based on minimization of the normalized cut criterion i.e. it maximizes the sum of the within-cluster similarity and minimizes the sum of between-cluster similarity. It is well adapted for clusters with a local point-to-point connectivity, so it does not depend on Gaussian distributions.

3. RESULTS

3.1. Selection of the optimal classifier

The implementation of PFB coupled with AOA allowed for the collection of 10 916 data points. For a given sampling period, when data was missing for at least for one parameter, the entire row was deleted. The time series was then composed of 7074 valid data points with 1573 data points for Leg 1, a total of 2192 for Leg 2 and 3309 for Leg 3. Missing data were

mainly caused by water flow interruptions when the water intake lifted above the waterline due to boat motion during bad weather conditions.

To highlight the added value of our approach in terms of clustering power, Fig. 2 illustrates clustering results on 3 different theoretical data series types: Fig. 2A consists of a simple case with 3 compact squares that are linearly separable. For this convex data, the 3 algorithms are all able to distinguish each square. Fig. 2B is 1 disk and 2 circles built from the same centre with an increasing radius. HC distinguishes only the convex disk with confusions with the 2 circles, and *k*-means does not succeed in separating the 3 forms. SC labels the data without any confusion. Fig. 2C has a more complex geometry with the data forming 3 close snake shapes. SC is the only method able to identify the 3 clusters without confusion.

For a large database, it is worth noting that a fast spectral clustering algorithm or a Nystrom approximation of eigenvectors for a large matrix can be performed, whereas the other methods, HC or *k*-means (also EM), require an unfeasibly high memory buffer capacity.

A theoretical labelling process was done to identify the different combinations of the 4 AOA groups that were observed during the cruises, to compare automatic clustering results obtained from the 3 proposed algorithms. Considering results presented in Figs. 3–5, the biologist defined, in a pragmatic manner, the most relevant and frequent combinations of AOA groups. Manual labelling was computed based on the combinatorial inclusion–exclusion principle for the 4 AOA groups. Each pairwise intersection, triple-wise intersection, quadruple-wise intersection and exclusive set (one group) of the set {BG, G, B, C} were associated with a specific label. Some sets were not present as a single group or as pairwise intersections with BG and are therefore not mentioned in Table 2. These combinations were also compared with a theoretical phytoplankton composition (at the Class taxonomic level; Table 1). Table 2 summarises these combinations and highlights the number of data points for each leg and each algae combination when considering the main patterns observed in Figs. 3–5. This labelling is called 'expert-based labelling' in the remainder of the paper.
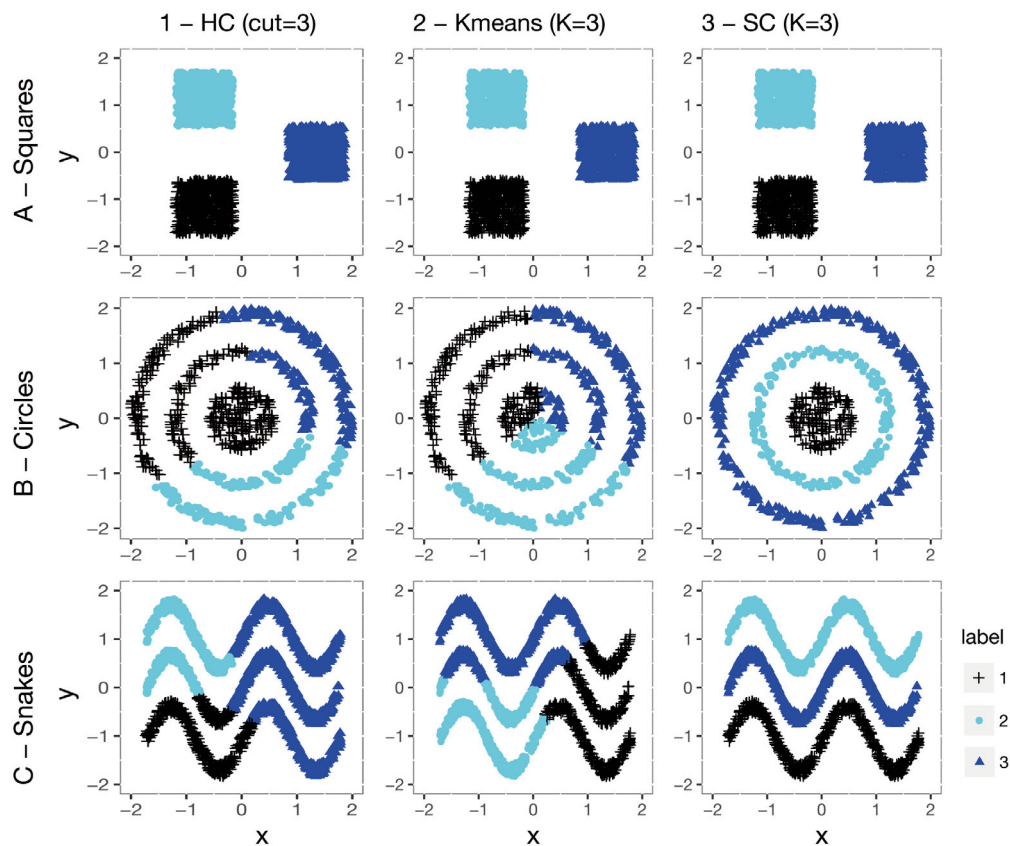

Fig. 2. Comparison of the performance of clustering algorithms (HC: hierarchical clustering; Kmeans: *k*-means; SC: spectral clustering) on artificial data structured differently in the classification space: (A) 3 compact square-shape clusters; (B) 3 clusters with 1 disk surrounded by 2 circles; (C) 3 streamers with point-to-point connectivity
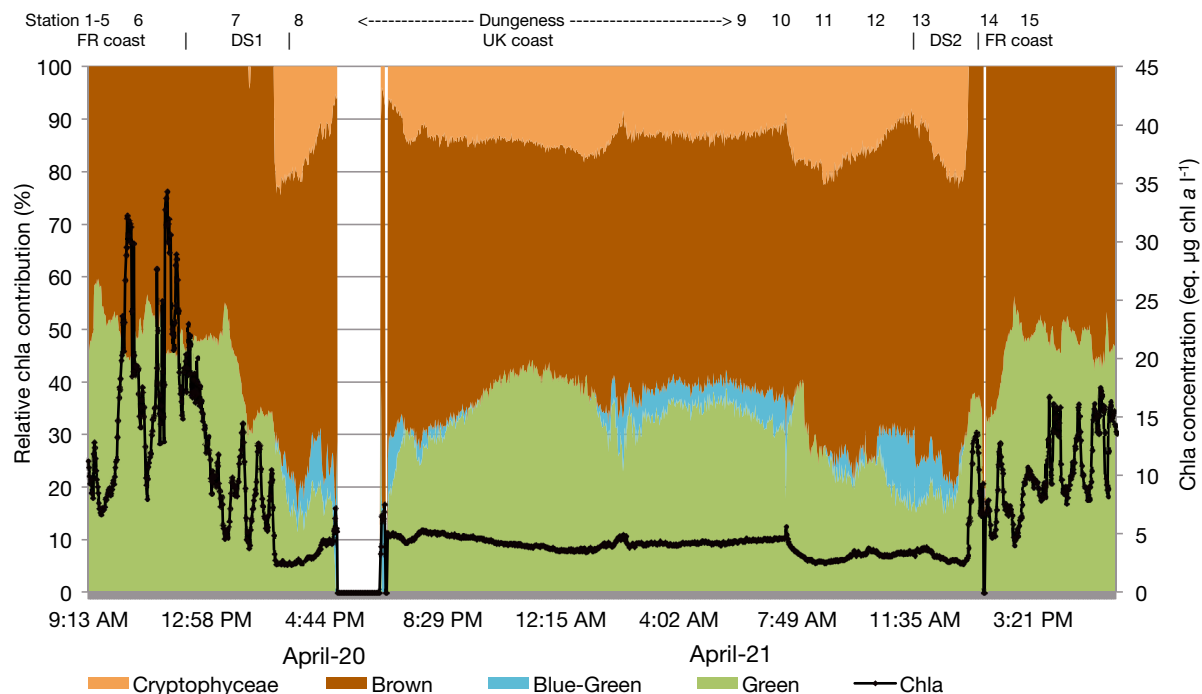
Fig. 3. Time series of chlorophyll *a* concentration as estimated by the Algae Online Analyser (AOA) (eq. µg chl *a* l⁻¹) (right axis) and relative contribution of each spectral group: green, blue-green, brown and Cryptophyceae in AOA parlance (%) (left axis) during Leg 1 (20–21 April 2012) (Stns 1–15 for conventional low resolution sampling). FR: French coast; DS: Dover Strait; UK: United Kingdom coast

Fig. 4. Time series of chlorophyll *a* concentration as estimated by the Algae Online Analyser (AOA) (eq. µg chl *a* l⁻¹) (right axis) and relative contribution of each spectral group: green, blue-green, brown and Cryptophyceae in AOA parlance (%) (left axis) during Leg 2 (27–29 April 2012) (Stns 34–53 for conventional low resolution sampling). FR: French coast; DS: Dover Strait; UK: United Kingdom coast
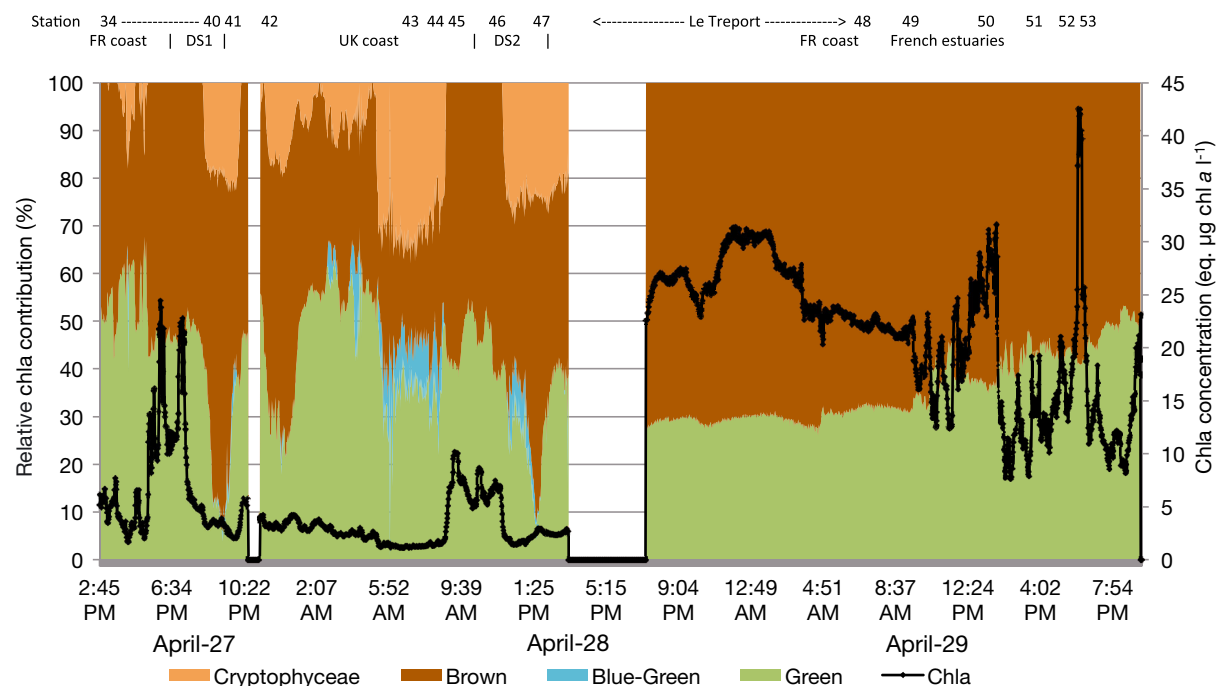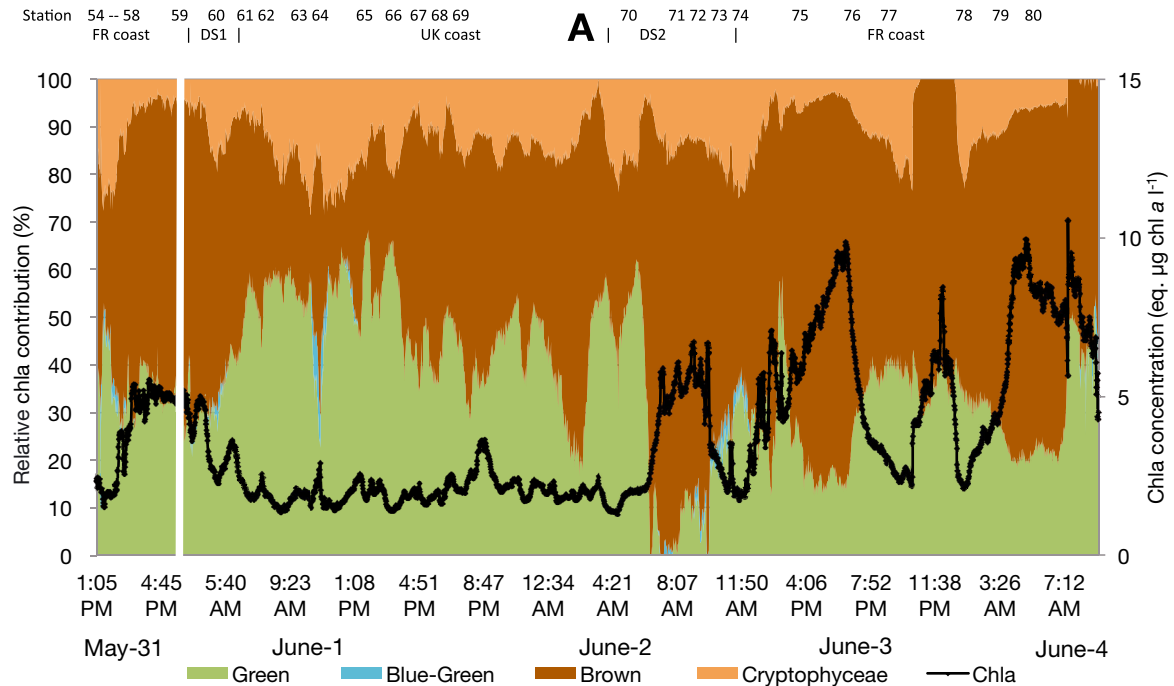
Fig. 5. Time series of chlorophyll *a* concentration as estimated by the Algae Online Analyser (AOA) (eq. µg chl *a* l^{-1}) (right axis) and relative contribution of each spectral group: green, blue-green, brown and Cryptophyceae in AOA parlance (%) (left axis) during Leg 3 (31 May–4 June 2012) (Stns 54–80 for conventional low resolution sampling). FR: French coast; DS: Dover Strait; UK: United Kingdom coast

Table 1. Relative abundance (%) of the main phytoplankton taxonomical classes identified during Leg 1 (Stns 8–15), Leg 2 (Stns 36–53) and Leg 3 (Stns 58–78) cruises in the eastern English Channel, from low resolution water samples, using inverted microscopy for identification. Bacillario.: Bacillariophyceae; Chloro.: Chlorophyceae; Crypto.: Cryptophyceae; Dictyo.: Dictyochophyceae; Dino.: Dinophyceae; Eugleno.: Euglenophyceae; Prasino.: Prasinophyceae; Prymnesio.: Prymnesiophyceae

Stn	Bacillario.	Chloro.	Crypto.	Dictyo.	Class Dino.	Eugleno.	Prasino.	Protista	Prymnesio.
Leg 1									
8	10.06	0	2.17	0	0.46	0	0	0.02	87.29
11	17.45	0	2.41	0	0.74	0	0	0.03	79.33
12	88.94	0	6.76	0	3.07	0.61	0	0.61	0
13	2.31	0	0.31	0	0.24	0	0	0.04	97.06
14	34.97	0	0	0	0.36	0	0	0.04	64.63
15	4.55	0	0.18	0	0.29	0	0	0.02	94.96
Leg 2									
36	4.96	0	0.11	0	0.5	0	0	0.11	94.32
40	90.53	0	9.18	0	0.2	0	0	0.04	0
42	83.83	0	11.84	0	4.34	0	0	0	0
44	3.13	0	0.74	0	0.16	0	0	0	95.87
48	12.27	0	0	0	0.33	0	0	0	87.4
49	4.44	0	0.12	0	0.2	0	0	0.08	95.17
50	4.27	0	0.24	0	0.37	0	0	0.13	94.99
51	2.71	0.03	0	0	0.09	0	0	0.05	97.12
52	3.75	0	0.68	0	0.71	0.02	0	0.1	94.73
53	2.01	0	0.01	0.01	0.05	0	0	0	97.93
Leg 3									
58	97.73	0	1.14	0	0	0	0	0	0
59	95.67	0	0	0	3.67	0	0	0.33	0
61	84.61	0	0	0	11.54	0	0	0.71	0
72	98.74	0	0	0	1.26	0	0	0	0
76	56.32	0	39.14	0.48	3.58	0	0	0.48	0
77	68.73	0	17.75	0	11.41	0	0	0	0
78	99.3	0	0	0	0.16	0	0	0	0.54

Table 2. Occurrence of several phytoplankton spectral group combinations (algal groups) and associated composition as identified by the Algae Online Analyser during each cruise (Legs 1 to 3) in the eastern English Channel. G: green algae; BG: blue-green algae; B: brown algae; C: Cryptophyceae

No. of algal groups	Composition	Occurrences Leg 1	Leg 2	Leg 3
1	G or BG or B or C	0	0	0
2	G–B	526	1394	587
	B–C	0	0	20
3	BG–G–C	0	0	10
	BG–B–C	0	0	23
	G–B–C	248	405	2260
4	All spectral groups	799	393	409

In the EEC areas travelled during the 3 legs, AOA detected at least 2 of the 4 phytoplankton groups. So for Legs 1 and 2, the number of clusters was set to 3, and then 5 to 6 for Leg 3. Table 3 presents the recognition rate (RR) and the adjusted Rand Index (ARI) for each classifier with each data set compared to the expert-based labelling. RR corresponds to the number of well-classified data points by the total amount of data (%); to assign the label, a class agreement by the maximum vote was computed. The RI computes a similarity measure between 2 clustering methods by considering all pairs of samples and counting the pairs that are assigned in the same or different clusters in the predicted and true clustering. ARI score [−1, +1] is a

corrected version of RI [0, 1], excluding the chance factor. The higher the positive ARI is, the closer the partitioning will be. A negative ARI value does not allow a confident similarity between the 2 clustering methods. ARI is an unsupervised score whereas RR is a score for supervised clustering. Here, classifiers have no training, so RR and ARI scores are not so well correlated. According to expert-based labelling, the classes are not homogenous, so a high RR score cannot be relevant. Indeed, it only means that the biggest class is well-recognized in contrast to the others (Table 3). For these experiments, the number of clusters is fixed to 3 or 6 classes. However, the number of relevant eigenvalues of the similarity matrix is more than 6 clusters, which means more information can be extracted from the spectral clustering. So, the strength of SC is not used at all in this part. For the following sections, as we wanted to analyse both hydrological and biological characteristics, only results from data set 4 were used.

3.2. Overview of general environmental conditions during the cruises

The first quarter of 2012 was characterized by low precipitation in the EEC (Meteo France data: https://donneespubliques.meteofrance.fr). In April, the oceanic rainfall disturbances occurred at regular intervals with precipitation values above normal. Strong wind events were measured, especially during the second half of April. Precipitation and wind conditions were typical in May. Abundant rainfall and strong SW sector winds occurred in June. From April to June, surface salinity fields were higher than the monthly averages, except in the Bay of Seine. Turbidity, which is expected to decline during this season, remained at relatively high levels (Cugier et al. 2005, www.umr-lops.fr/Donnees/MARC).

Cruises took place during periods of low or high tidal ranges, respectively for Leg 2 and then for Legs 1 and 3.

Statistical summaries for each parameter and for each leg are provided in Table 4. Salinity fluctuated between 34.0 and 34.6 during all legs. Water temperature was 9.0°C during the first 2 legs, but reached 13.0°C during Leg 3. Oxygen saturation was always

Table 3. Comparison scores (RR: recognition rate; RI: Rand Index; ARI: adjusted Rand Index) between partitions from automatic clustering (HC: hierarchical clustering; KM: *k*-means; SC: spectral clustering) and phytoplankton composition labelling (2, 3 or 4 algal groups, and then for the 6 detailed groups) for each cruise (Legs 1 to 3) and each data set

Data set	RR HC	KM	SC	RI HC	KM	SC	ARI HC	KM	SC
Leg 1									
1	0.82	0.59	0.37	0.82	0.79	0.78	0.61	0.55	0.57
2	0.61	0.01	0.00	0.78	0.77	0.80	0.54	0.52	0.61
3	0.37	0.05	0.54	0.70	0.73	0.80	0.40	0.46	0.61
4	0.69	0.00	0.00	0.78	0.78	0.78	0.56	0.56	0.57
Leg 2									
1	0.11	0.34	0.53	0.74	0.7	0.61	0.47	0.39	0.21
2	0.35	0.43	0.40	0.67	0.63	0.45	0.34	0.26	−0.08
3	0.07	0.32	0.19	0.67	0.66	0.43	0.33	0.31	−0.09
4	0.11	0.60	0.30	0.66	0.71	0.42	0.31	0.42	−0.13
Leg 3									
1	0.22	0.00	0.00	0.62	0.64	0.63	0.26	0.29	0.27
2	0.09	0.01	0.16	0.60	0.61	0.60	0.22	0.23	0.22
3	0.21	0.09	0.10	0.54	0.55	0.53	0.10	0.11	0.08
4	0.04	0.29	0.08	0.64	0.61	0.58	0.30	0.23	0.18

Table 4. Statistical summaries for each parameter from the Pocket Ferry Box (PFB) (Salinity; Temp.: temperature; Oxygen: oxygen concentration; Sat.: saturation) and from the spectral fluorometer (Algae Online Analyser, AOA) (G: green algae; BG: blue-green algae; B: brown algae; C: Cryptophyceae; Chl *a:* chlorophyll *a*) (eq. µg chl *a* l^{-1}) and for each cruise (Legs 1 to 3)

Statistic	Salinity (PSU)	Temp. (°C)	Oxygen (mg l^{-1})	Sat. (%)	(eq. µg chl *a* l^{-1})				
					G	BG	B	C	Chl *a*
Leg 1									
Minimum	33.3	8.51	7.68	84.8	0.00	0.00	0.85	0.00	2.37
1st Quartile	34.5	8.82	8.35	92.2	1.01	0.00	1.80	0.00	3.60
Median	34.8	8.94	8.48	93.3	1.49	0.01	2.19	0.49	4.35
Mean	34.6	8.94	8.57	94.2	2.72	0.11	3.60	0.36	6.78
3rd Quartile	34.8	9.09	8.67	95.1	4.06	0.19	4.50	0.57	8.54
Maximum	35.0	9.48	9.66	106.5	15.60	1.52	18.69	3.91	34.30
na	21	21	23	21	448	448	448	448	448
Leg 2									
Minimum	33.3	9.04	8.03	89.4	0.00	0.00	0.17	0.00	1.15
1st Quartile	33.7	9.55	8.35	93.0	1.11	0.00	1.46	0.00	2.91
Median	33.9	9.73	8.51	94.4	3.14	0.00	4.09	0.00	7.02
Mean	34.0	9.71	8.53	94.8	4.07	0.03	6.86	0.14	11.10
3rd Quartile	34.4	9.90	8.77	96.9	6.87	0.00	12.62	0.29	20.70
Maximum	34.5	10.25	9.15	101.4	17.54	8.81	25.01	1.10	42.50
na	238	238	242	238	905	905	905	905	905
Leg 3									
Minimum	31.6	11.6	7.33	88.5	0.00	0.00	0.19	0.00	1.32
1st Quartile	33.8	12.9	8.13	97.4	0.88	0.00	0.89	0.14	2.08
Median	34.4	13.5	8.67	103.5	1.17	0.00	2.14	0.29	3.73
Mean	34.2	13.5	8.60	103.3	1.48	0.01	2.91	0.28	4.69
3rd Quartile	34.8	14.1	8.99	108.0	1.65	0.00	4.48	0.41	6.87
Maximum	35.0	15.8	10.40	122.5	7.35	0.80	9.89	1.10	15.80
na	1423	1423	1434	1423	2276	2276	2276	2276	2276

higher than 94%, rising to 122.5% during Leg 3. Oxygen concentration varied from 7.33 to 10.40 mg l^{-1}, with the highest values occurring during Leg 3.

3.3. Phytoplankton biomass, composition and spectral groups

Total phytoplankton biomass (i.e. chl *a*) as measured by AOA varied from 1.15 to 42.50 eq. µg chl *a* l^{-1} between late April and early June (between Legs 1 and 3) in the studied area. Maximum concentrations were measured during Leg 2 (median: 7.02 eq. µg chl *a* l^{-1} during Leg 2; respectively 4.35 and 3.73 eq. µg chl *a* l^{-1} during Legs 1 and 3).

Microscopic analyses of discrete samples collected during the different cruises (Table 1) highlighted dominance in the phytoplankton community of the Prymnesiophyceae *Phaeocystis globosa* (maximum 1.0×10^7 cells l^{-1}, varying according to cruise and almost exclusively on the French side of the Channel), and/or Bacillariophyceae (maximum 3.0×10^6 cells l^{-1}), with smaller, varying amounts of Cryptophyceae (maximum 4.3×10^4 cells l^{-1}) and dinoflagellates

(maximum 3.5×10^4 cells l^{-1}). *P. globosa* contribution to total cell numbers reached 98% in some areas, but the species was also absent in other areas, mainly in the central Channel and on the UK side of the Channel. Although some dinoflagellates (*Gymnodinium* sp., *Gyrodinium spirale*, *Katodinium* sp., *Protoperidinium* sp., *Torodinium* sp.) were identified, they contributed a maximum of 3% of total cell numbers in samples (maximum contribution of 3% for *Gymnodinium* sp., 1.3% for *Katodinium* sp., while the others remained below 1%).

Legs 1 and 2 (Figs. 3 & 4) occurred during the main bloom of *P. globosa*, whereas this bloom was on the wane during Leg 3 (Fig. 5).

During Legs 1 and 2, French waters were generally characterized by higher chl *a* concentrations (maximum of 34.30 and 42.50 eq. µg chl *a* l^{-1}, respectively). They showed a high spatial variability of chl *a* concentration within a given area dominated by *P. globosa*, i.e. where the B/G group signal ratio was approximately 50/50 (Figs. 3 & 4).

Temporal variability was also important. For a given French water mass, chl *a* concentration was highly variable from day to day: for example, a maxi-

mum chl *a* concentration of 34.29 eq. µg chl *a* l^{-1} was measured on 20 April compared to a maximum of 17.49 eq. µg chl *a* l^{-1} measured on 21 April (Fig. 3). The chl *a* concentration near Boulogne-sur-Mer was higher (~10 eq. µg chl *a* l^{-1}) at the end of Leg 2 (29 April) compared to the beginning (~5 eq. µg chl *a* l^{-1}) (27 April). Moreover, our high resolution sampling strategy allowed us to highlight changes in the relative proportion of spectral groups for a given chl *a* concentration and change in chl *a* concentration for a given proportion of spectral groups moving with the tide (fixed point offshore from Dungeness, UK, on 20–21 April, offshore from Le Tréport, Fr, on 28–29 April) (Fig. 4).

During Leg 3, chl *a* concentration was lower (maximum of 15.80 eq. µg chl *a* l^{-1}) and the characteristic *P. globosa* signal was less dominant. The highest chl *a* concentrations were observed in areas under freshwater influence (offshore from the 3 main rivers on the EEC: Canche, Authie, Somme or/and under the influence of the Seine River) (during Leg 1: DPM 49 to 51; during Leg 2: DPM 70 to 80).

UK waters had lower chl *a* concentrations (<5 eq. µg chl *a* l^{-1}) with a minor contribution of Cryptophyceae and BG spectral groups during Legs 1 and 2 (Figs. 3 & 4).

3.4. Unsupervised classification

In order to highlight the added value of our methodology, we now focus on Leg 2 data, which contains the lowest number of NA values (2192 complete AOA–PFB multi-parameter measurements), proposes interesting variabilities to study and reflects characteristics of the English Channel water masses when partially dominated by *P. globosa* blooms. The number of cluster *k* is now automatically computed. With the eigen gap method, which looks for the maximum consecutive gap between sorted eigenvalues, *k* is equal to 85 with eigenvalues higher than 0.78. A proposal with *k* = 85 corresponds to an optimized partitioning, and consequently is equivalent to an optimization of the extraction of the information contained in the HF data series. Nevertheless, for the time being, expertise is not sufficiently developed to further interpret, from an ecological point of view, such an amount of information. Indeed, each of the 85 groups is relative to a specific environmental state characterized by a specific geometrical arrangement in spectral space. By restricting this criterion with a condition on eigenvalues (λ) greater than 0.98, *k* is equal to 18, and for λ > 0.99, *k* is equal to 6 (Fig. 6). Cluster 0 corresponds to part of the cruise when
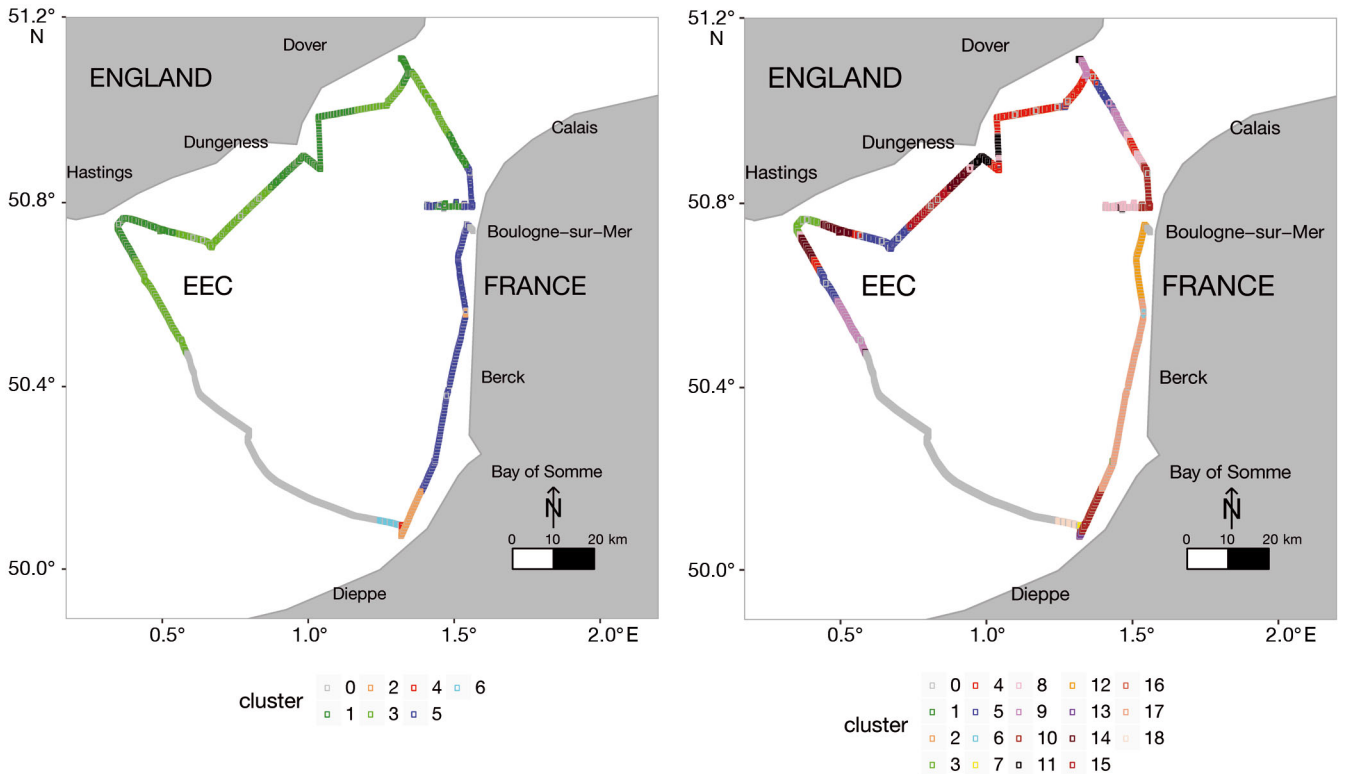

Fig. 6. Temporal and spatial view of the resulting clusters of Leg 2 by spectral clustering for *k* = 6 (left hand side) and *k* = 18 clusters (right hand side) (colours between panels are not related)

some PFB or AOA data were missing; therefore, these data were not used during the classification process.

For $k = 6$, four clusters highlight the B and G AOA spectral groups in variable proportions (Clusters 2, 4, 5 and 6) (Fig. 7). Cluster 1 shows a small proportion of the Cryptophyceae signal, whereas the last cluster, 3, contains all 4 AOA spectral groups with a relatively high proportion of the Cryptophyceae signal, to the detriment of the G signal. Only Clusters 1 and 3 are characterized by the BG spectral group (Fig. 7).

3.5. Biotic and abiotic interactions

After unsupervised classification, it is possible to highlight the main statistical characteristics of the studied parameters for each cluster. With $k = 6$, statistics are summarised using boxplots in Fig. 8.
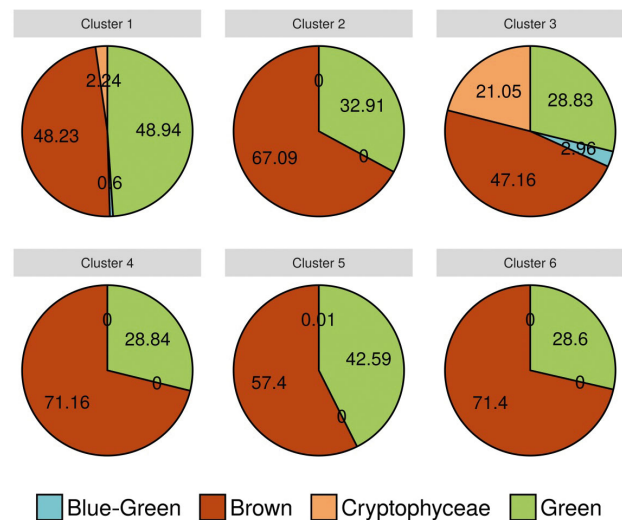
Fig. 7. Relative proportion (%) of each of the 4 Algae Online Analyser spectral groups (blue-green, brown, Cryptophyceae and green) for each cluster ($k = 6$) during Leg 2
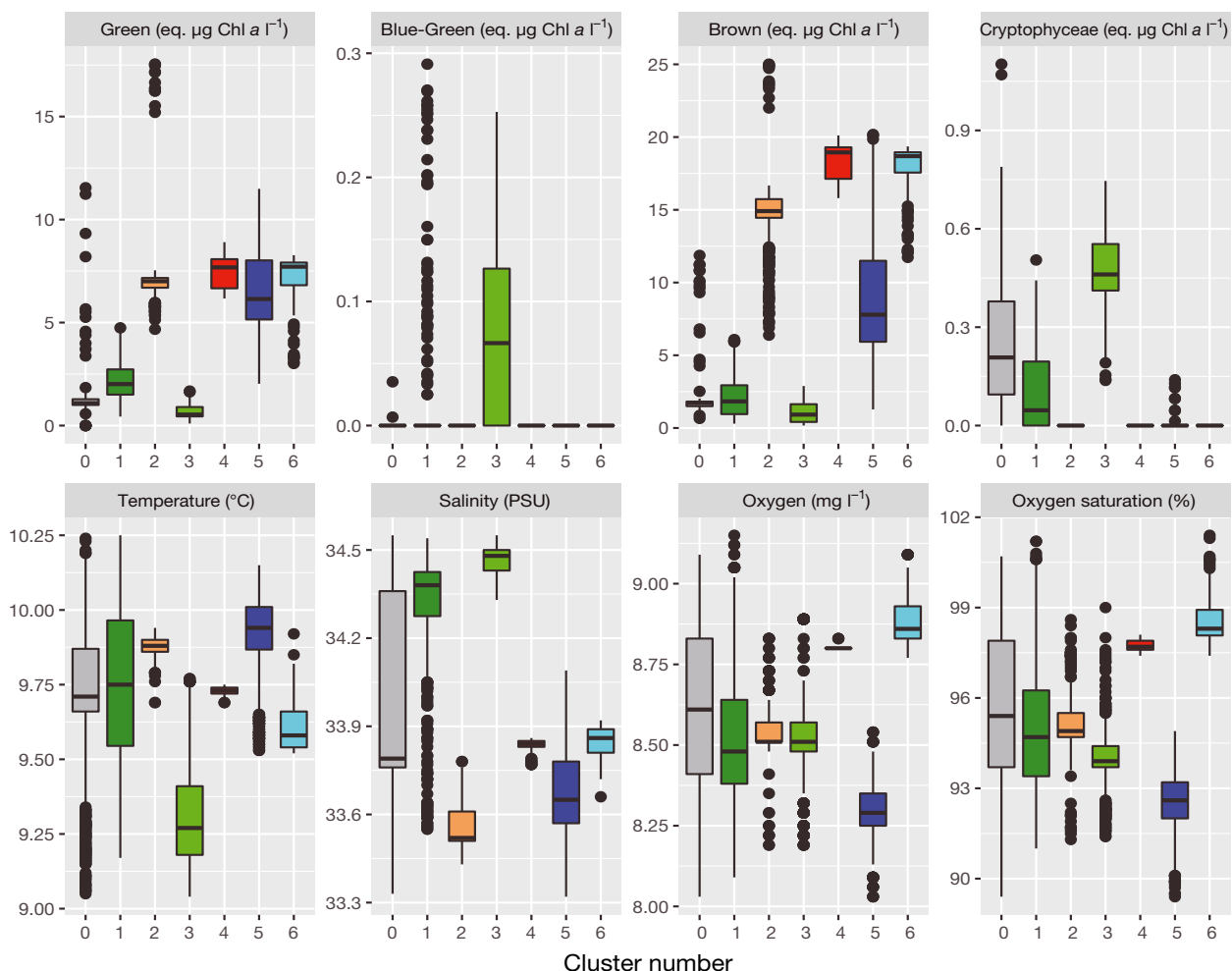
Fig. 8. Parameters measured during Leg 2 (27–29 April 2012) and for each cluster according to the spectral classification ($k = 6$). Bar: median; box: interquartile range; whiskers: 10th–90th percentiles; dots: outliers

Table 5. Pearson correlation coefficients between Pocket Ferry Box (PFB)/ Algae Online Analyser (AOA) parameters and clusters for $k = 6$ and for Leg 2

Parameter	Cluster					
	1	2	3	4	5	6
AOA						
Green	−0.39	0.37	−0.64	0.23	0.44	0.23
Blue Green	0.04	−0.16	0.44	−0.09	−0.24	−0.10
Brown	−0.44	0.46	−0.52	0.40	0.15	0.40
Cryptophyceae	−0.10	−0.26	0.91	−0.15	−0.40	−0.16
PFB						
Temperature	0.09	0.24	−0.73	0.02	0.46	−0.09
Salinity	0.42	−0.46	0.65	−0.10	−0.54	−0.10
Oxygen concentration	0.04	0.08	0.09	0.31	−0.56	0.43
Oxygen saturation	0.15	0.12	−0.12	0.34	−0.53	0.46

The calculation of the correlation matrix makes it possible to prioritize the contribution of the physico-chemical and biological parameters to the structuring of each group, subsequently for each water mass, and for each environmental state (Table 5). Partitioning clearly reveals an environmental difference in terms of the spectral signature of phytoplankton between the French and English coasts, characterized by Clusters 2, 4, 5, 6 and 1 and 3, respectively (Figs. 6 & 7). Occurrence of the BG group characterizes Cluster 3 (2.96%). In this cluster, the Cryptophyceae signal is also very important (21.05%). More generally, on the French side of the EEC, it is the proportion between the B and G signals that makes the difference. Indeed, Cluster 5 represents water masses with a high proportion of *P. globosa*; almost half of the signal is distributed in the B and G groups. In these water masses, salinity is rather low (median 33.6), oxygen concentration (median 8.29 mg l^{-1}) and oxygen saturation (92.6%) are the lowest, and temperature is the highest (median 9.94°C). Salinity is maximum for UK water masses (Clusters 1 and 3). Water temperature is the highest for Cluster 1 (median 9.75°C) compared to Cluster 3 (median 9.27°C). Clusters 2 and 4 highlight changes related to the tidal cycle during the Eulerian sampling period offshore from Le Tréport (Fr) on 28–29 April 2012 (> DPM 47 and < DPM 48; Fig. 4). Cluster 6, with a spectral signature close to that of Clusters 2 and 4 (approx. 50/50 G/B), corresponds to an interim group between water masses dominated by *P. globosa* (95–98% of the total phytoplankton abundance) in the northern part of the EEC and other water masses with a higher diatom proportion in the southern area (*P. globosa* < 90%). In Cluster 6, salinity is low (median 33.5), while oxygen concentration (median 8.54 mg l^{-1}) and oxygen saturation (95%) are inter-mediate compared to that of the other clusters.

In order to have a synoptic view of the main structuration of spectral phytoplankton groups in the EEC, we merged the data sets of Legs 1, 2 and 3. According to the same methodology, we obtained $k = 291$ in unsupervised mode. With criteria $\lambda > 0.996$ and $\lambda > 0.990$, we generated, respectively, $k = 9$ and $k = 33$ (Fig. 9). Thus, these clusters make it possible to integrate both the spatial and temporal variability of the phytoplankton spectral groups and the parameters measured by the PFB/AOA. The difference in the number of groups identified between the water masses of the English and French waters is clear. UK waters are more heterogeneous than those on the French side. The offshore water masses are relatively well individualized.

4. DISCUSSION

The value of our approach is to propose a concrete and complete demonstration of the possibilities offered by a sampling strategy based on high frequency (<1 h) automated measurement systems with a data processing step based on an optimal classifier in order to extract the best available information for water quality assessment issues. A multiple-fixed-wavelength spectral fluorometer (i.e. AOA) coupled with a PFB was implemented as a means of quantifying phytoplankton biomass and community structure as well as main physicochemical characteristics. The implementation of the AOA is commonly presented as a useful tool for monitoring phytoplankton community composition only if one implements frequent calibration by discrete sample collection, calibration with species representative of the region of interest and additional HPLC analysis (Richardson et al. 2010, MacIntyre et al. 2010, Houliez et al. 2012). In this paper, we proposed a pragmatic solution that does not require labour-intensive analyses. Indeed, we highlight that even if we used the AOA in its original configuration (i.e. with a minimum amount of time and effort devoted to calibration procedures), it is possible to change the way we track phytoplankton, including HABs.

The high-resolution along-route approach implemented during our cruises in the EEC in 2012 uncovered a series of results that clearly confirmed the
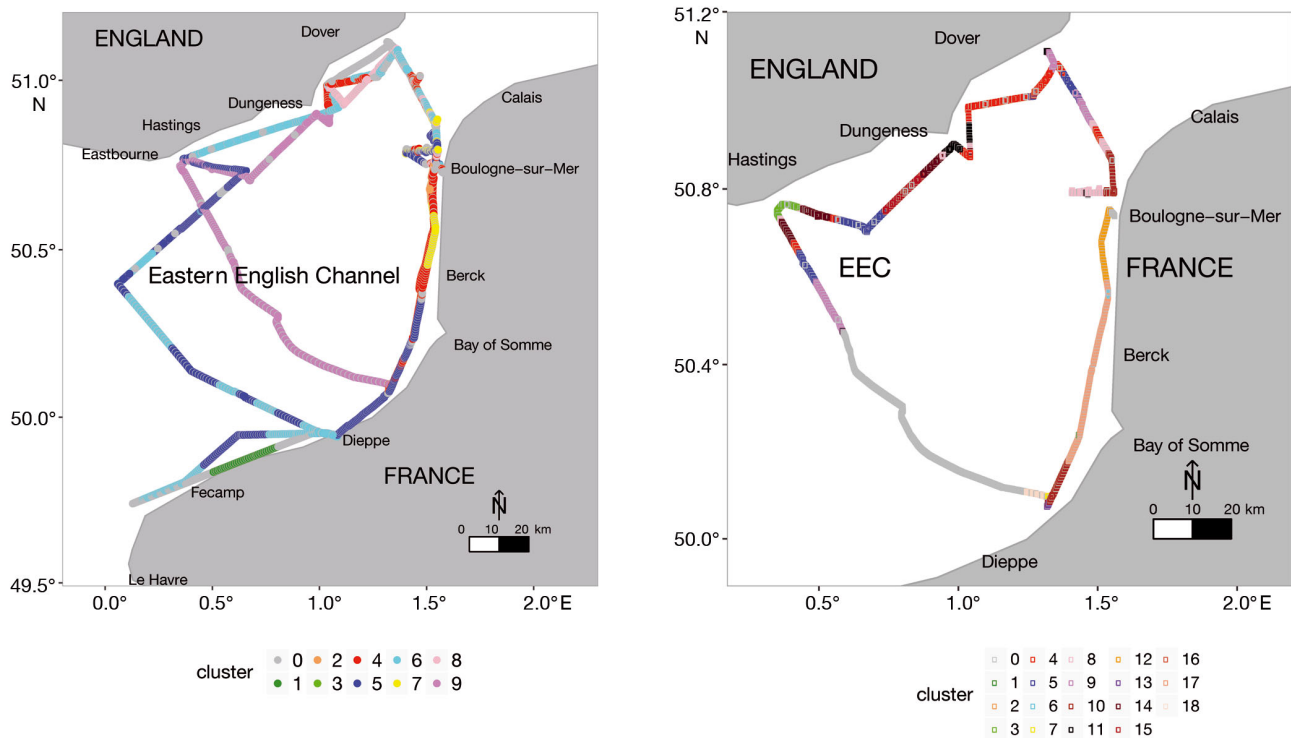
Fig. 9. Spatialized projection of the 9 (left hand side) and 33 (right hand side) clusters defined by spectral clustering based on high frequency data series collected during Legs 1, 2 and 3 (20 April–4 June 2012) (colours between panels are not related)

added value of taking continuous measurements during cruises. It provides a more detailed view of the spatial distribution (and to a lesser extent in our case study, the temporal resolution) of the phytoplankton spectral groups and associated water mass characteristics in the EEC. The originality of our approach is that we propose an unsupervised spectral classification method to extract as much information as possible from the high-resolution multi-parameter database.

The advantage of using an unsupervised approach by spectral classification is to remove any uncertainty about the distribution of data, as opposed to conventional approaches such as expectation–maximization or *k*-means (based on a Gaussian distribution hypothesis), or hierarchical classification (HC implies the definition of a cutting distance). Otherwise, HC relies on a strong initial partitioning choice that may bias or skew the results, while spectral partitioning relies on a global approach.

The algorithms presented here are constrained to a complete database without missing data.

Based on recent works from Phan et al. (2017), we could expect to implement in the very short-term multiple imputation and data completion methods using dynamic time-warping algorithms to enhance

our database in order to improve environmental state/cluster discrimination.

This approach confirms that it is possible to obtain a near real-time synoptic view of the distribution of 4 phytoplankton spectral groups (preliminary taxonomic approach) and of the main characteristics of the associated water masses. Moreover, in slightly delayed time, it is also possible to define environmental states, characterized by a specific combination of physicochemical and biological parameters, when implementing an unsupervised spectral classification approach.

During the 3 cruises between late April 2012 and early June 2012, implementation of the PFB coupled with the spectral fluorometer AOA allowed us to characterize and quantify the high spatial and temporal variability of the phytoplankton biomass and associated spectral groups.

Characteristics of Cluster 3 (UK offshore and central English Channel water masses with high salinity, low temperature, intermediate oxygen concentration and saturation) are conditioned by the Atlantic waters which are channeled north-eastward (Salomon & Breton 1993).

During transport towards the North Sea, water masses are modified by local riverine inputs, espe-

cially along the French coast under the influence of the Seine River (and to a lesser extent the Somme, Canche and Authie rivers). Nutrient concentration (nitrate, phosphate and silicate) is higher in these coastal waters, which are separated from the offshore water bodies by a frontal structure which evolves with the tides, mainly in response to the neap-tide/spring tide cycle (Brylinski et al.1991). This contributes to the creation of favourable environmental conditions for phytoplankton blooms, and consequently, for *Phaeocystis globosa* (Cluster 5). This result is consistent with other observations made in the same area (Lefebvre et al. 2011, Bonato et al. 2015) or in the neighbouring Belgian coastal zone (Muylaert et al. 2006). These water masses (Cluster 5) correspond to a sensitive area regarding eutrophication issues. The most visible indirect effect of *P. globosa* blooms in these water bodies is a decrease in oxygen concentration and saturation. Indeed, oxygen saturation over 100% generally mimics the phytoplankton biomass pattern due to primary production, but in some cases, under-saturation occurs where bacterial consumption is associated with decaying algal blooms or degradation of excessive amounts of foam, as in the case of the wane of a *P. globosa* bloom.

The absence of nutrient data (because of analytical problems) does not undermine this interpretation, since it is the winter nutrient concentrations and the dynamics of their use during the spring diatom bloom that will explain the proliferation (or not) of *P. globosa* under adequate weather and hydrodynamic conditions. Using complementary low resolution data from Ifremer's monitoring programmes REPHY and SRN (Lefebvre et al. 2011, SRN 2017), we observed maximum dissolved inorganic nitrogen, phosphate and silicate concentrations of 34.43, 0.94 and 17.0 µmol l^{-1}, respectively offshore from Boulogne-sur-Mer, and 42.28, 1.03 and 24.4 µmol l^{-1}, respectively in the coastal zone of the Bay of Somme. These winter concentrations are consistent with what is usually observed in the area, and are compatible with the development of an intermediate magnitude spring bloom of *P. globosa* (Lefebvre et al. 2011, Hernández-Fariñas et al. 2014).

Clusters 2 and 4 make it possible to appreciate the variability of the biotic and abiotic conditions during a tidal cycle (since the boat was at anchor because of bad weather conditions, i.e. as in an Eulerian monitoring strategy).

Cluster 6 shows a relative proportion of the spectral groups that are significantly different from that of the more northern watersheds (Cluster 5), in accordance with a previous hypothesis based on a low resolution sampling strategy (Brunet et al. 1996) indicating different structuring of the phytoplankton community between the south and north region of the Bay of Somme.

Without any strong riverine input, UK coastal water masses and the central EEC showed similar characteristics, highlighted in Cluster 1.

The high variability of phytoplankton biomass is relative to a patchy distribution of phytoplankton generally observed at scales from centimetres up to several kilometres (Seuront 2005); that author highlighted that in French coastal waters in the northern part of the EEC (under the influence of a tidal front and considering the tidal and wind-driven alongshore transport), it is not surprising to observe patchiness at a kilometric scale. Changes in phytoplankton biomass and the relative proportion of spectral groups along the French coast at the end of Legs 2 and 3 are related to a change in local conditions in response to riverine nutrient inputs from the Somme and Canche rivers.

When merging data from Legs 1, 2 and 3, our methodology allows automatic separation of 33 clusters. Consequently, in spring 2012, the EEC could be considered as a complex environment, with 33 combinations in the spectral space of the main variables we measured. Thanks to improving technologies and reduced costs in recent years, we moved from monitoring strategies consisting of explaining 'how to measure such or such variable using new technologies' to a new approach where the main limitation seems to be our own capacity, i.e. expertise and interpretation of huge amounts of data.

In situ measurements compared to reference discrete samples during the cruises highlighted that *P. globosa* was classified by AOA as a mixture of G and B algae. In this case, the signal was characterized by a relative proportion of approximately 50% G and 50% B signals. In the EEC, Houliez et al. (2012) noted that, when using the original fingerprints of a Fluoro-Probe (the same technology as in the AOA), the classification was dominated by the B signal all year, except during April–May (the *P. globosa* bloom period) where the contribution of the G signal was close to that of the B. Consequently, in our study, and considering the absence of G algae in the sampled area, the first idea was to use a new specific fingerprint generated from cultures of *P. globosa*, and so a specific combination of spectral groups—*P. globosa*, BG, B and Cryptophyceae—rather than using the original fingerprints as proposed by the manufacturer. The core issue is that with our AOA, the library

of fingerprints is limited to 4 fingerprints, whereas we have several dozens of phytoplankton species in the area. Therefore, it is not possible to include the fingerprints from all taxonomic phytoplankton groups found in the studied area (Restriction 1). In addition, we know from our own tests and from the literature (Seppälä & Olli 2008, Houliez et al. 2012, Escoffier et al. 2015) that if we force AOA to attribute part of the signal to the *P. globosa* group, it will lead to major errors when this species is absent. Indeed, the relative abundance of all groups must equal 100 % because phytoplankton fluorescence responses are assigned to 1 of the only 4 available spectral groups based on similarity of spectral signatures (Restriction 2). As we wanted to sample a large area within a relatively small time window, we anticipated that the phytoplankton community will not only be dominated by *P. globosa*, and that the use of co-varying fingerprints is not allowed (Restriction 3). The specific fingerprints derived from cultures do not represent variability found *in situ* (diversity in pigments, and thus spectra; different physiological responses to different environmental conditions) (Restriction 4). Finally, results are not comparable between different AOAs and for a given AOA, results are not comparable before and after factory calibration when servicing and maintenance are needed (Restriction 5).

Considering these restrictions, we decided not to use field-derived norm spectra as proposed by Harrison et al. (2016). As our objective was to propose an analytical rapid procedure based on the KISS principle (Keep It Smart and Simple), we want to (1) make the implementation of the PFB/AOA devices as simple and as fast as possible for non-specialists (i.e. without any fine-tuning of the devices and without laborious experiments with cultures to generate new specific and regionalized fingerprints); (2) rapidly generate data on the spatio-temporal dynamics of phytoplankton spectral groups without any particular *a priori* knowledge of the biological characteristics of the studied area, and without any particular expertise or specialized training required of the user; and (3) propose a data processing procedure based as closely as possible on the raw data in order to limit expert judgment and a long-lasting data pre-processing stage (i.e. an adjustment of the fingerprints to specific taxa in large areas with a high biodiversity/heterogeneity would require a post hoc modification of the fingerprint library configuration, only possible after confirmation of the occurrence of such taxa via a time-consuming microscopic enumeration).

We had then to compromise between taxonomic resolution and user-friendly implementation. Our choice was to work with original fingerprints in order to (1) be able to tackle changes in phytoplankton structure with a high spatio-temporal resolution, such as a screening procedure before implementation of another more directed sampling strategy (i.e. discrete samples for conventional microscope-based analysis, or flow cytometry analysis); and (2) make results from different seasons and different areas as comparable as possible (consequently with a potentially high diversity of spectral fingerprints incompatible with the AOA limited library as proposed for the time being by the manufacturer).

Finally, despite a rather limited taxonomic resolution to define the structure of phytoplankton communities, it appears that our results are very close to those of Bonato et al. (2015). Implementing scanning flow cytometry (SFCM), and using a large matrix (sampling locations × optical properties of cells), they were able to define 4 types of water masses during Leg 2 using hierarchical classification (Bray-Curtis dissimilarity index based on Ward distance). It seems that the differentiation of water masses according to the phytoplankton structure of the communities is quite comparable when using either cell level (SFCM) results discriminated with a common hierarchical classification, or bulk level (AOA) results analysed with an optimized spectral classification methodology. The use of one or the other approach would only be justified because of the need for taxonomic resolution and acquisition of cell structural parameters. Apart from the scientific objectives, the major differences lie in the equipment cost (SFCM > AOA by a factor of 6) and in the time required to extract the information contained in the data series (SFCM > AOA by a factor of 60, i.e. 30 vs. 0.5 d of data processing over a 1 mo duration cruise, with a SFCM sampling frequency of 10 min vs. an AOA frequency of 1 min).

Focusing on *P. globosa*, it is noteworthy that whereas species cannot be differentiated in an automatic way (considering Restrictions 1 and 2 in particular), an increasing proportion of the G signal associated with the B one in the French part of the EEC should be considered as a possible early signal of the initiation phase of a *Phaeocystis* bloom.

With only 3 cruises focalized on a small time window, i.e. the spring bloom, we were able to provide new results regarding the spatial and temporal distribution of phytoplankton spectral groups and associated water properties, and their variability during a *P. globosa* bloom, i.e. during an HAB. All these variabilities would be difficult to detect by other low resolution routine monitoring parameters alone,

or are too prohibitively expensive to be undertaken as part of a regular monitoring program using conventional approaches (water samples using Niskin bottles, on-board filtrations, laborious laboratory analyses). Consequently, generalization of this kind of high resolution approach on larger data sets (i.e. on extended spatial and temporal monitoring periods such as the perennial Ferry Box line) represents a significant new data source to complement existing low resolution, conventional monitoring programs. It is important to note that the idea was not to propose an alternative to conventional approaches, but to develop complementary tools and to gain experience, so other researchers could make a choice considering the advantages and drawbacks according to their scientific objectives. Such a strategy has already been recommended in other contrasting environments (i.e. differing by their main hydrodynamical and hydrobiological characteristics) by, for example, Richardson et al. (2010) for the Neuse River estuary (Georgetown, SC, USA), Aiken et al. (2011) for the Chilean Fjords and Catherine et al. (2012) for lakes and reservoirs.

To summarise, our approach enabled us to (1) propose a well-fitted and adaptive sampling strategy in near real-time during a cruise (monitoring objectives are better defined if geographical boundaries are set; sampling can be enhanced for areas with high environmental gradients and when high variability exists for a given parameter); (2) implement an early warning system to track changes induced by HAB events; (3) more generally, to improve knowledge on the spatio-temporal distribution of phytoplankton communities and also knowledge of pelagic habitats; and (4) identify controlling factors and direct and indirect effects of massive bloom events (including HABs), with the proviso of adding supplementary supporting factors while monitoring.

This optimal differentiation of environmental states/ water masses by spectral clustering is primarily significant considering the need for hidden Markov modelling (HMM) as proposed by Rousseeuw et al. (2015). HMM allows one to characterize the environmental states and to model their dynamics (emission probabilities and transition probabilities between states) but also to predict new or future states.

Species composition and the relative abundance of different algal groups in a community are the fundamental determinants of aquatic ecosystem structures and functions. Alterations in phytoplankton community composition can lead to negative ecological impacts on entire ecosystems; for example, eutrophication, HABs and hypoxia (See et al. 2005). Consequently, some criteria dedicated to defining the GES

of the MSFD are related to phytoplankton (D5C2: chl *a* in the water column; D5C3: HABs) and pelagic habitats (D1C6: pelagic broad habitat types), and therefore need to be integrated into the monitoring program. At this stage of MSFD implementation, the effect of excessive nutrient inputs would be primarily measured by changes in the phytoplankton biomass, hence its proxy, the chl *a* concentration, which seems insufficient when taking into account the complexity of the phytoplankton response to changes in environmental pressures (Lefebvre et al. 2011, Hernández-Fariñas et al. 2014, this study). Moreover, Alvarez-Fernandez & Riegman (2014) postulated that in an environment evolving towards a limitation of phytoplankton growth by nutrients, and subject to increasing temperatures and changing light conditions because of global changes, the use of chl *a* concentration measurements as an indicator of phytoplankton biomass should be considered with caution. Such a simplification of the assessment using change in phytoplankton biomass as the only indicator of change in the phytoplankton community could lead to 2 situations that are counterproductive to one another: (1) the assessment may lead to the conclusion that no change in biomass has occurred, and consequently no environmental measures would be undertaken (no recommendation for additional requirements for reducing inputs). However, for a given stable biomass, the structure of the community and also the ratio of chl *a* to carbon concentration can be significantly different. In our study, we showed that changes can be favourable to the proliferation of potentially harmful species (*Phaeocystis* sp.). Lack of management actions could lead to deterioration in this situation. (2) The assessment may lead to the implementation of additional measures to reduce nutrient inputs when an increase in phytoplankton biomass is observed. If this increase in biomass is only linked to the development of 'forage' phytoplankton, then these measures will lead to limiting energy transfers within the trophic network, which may be beneficial for higher trophic levels by offsetting other pressures. In this case, management action may inhibit an improvement in the situation.

The strategy of high-resolution measures of environmental parameters presented here is in line with the recommendations and work undertaken by the FerryBox community (www.ferrybox.org) in the Baltic Sea (Rantajärvi 2003), which was then extended to other European marine areas thanks to projects such as FP5 FerryBox, FP7 Protool then H2020 JERICO-Next (Petersen et al. 2008, Thyssen et al. 2015, www. jerico-ri.eu/). On the basis of the experience gained

from this proof-of-concept in the EEC, and considering the promise to improve scientific knowledge through the use of optimized numerical data processing tools, we have compiled guidance for consideration in the development of a new Ferry Box line across the EEC.

5. CONCLUSIONS AND PERSPECTIVES

The implementation of high resolution *in situ* analysis using a PFB coupled with a spectral fluorometer and associated with an optimized unsupervised spectral classifier for data processing allowed us to draw an environmental state-specific synoptic view of the main hydrological parameters, phytoplankton biomass and spectral groups, in the EEC while a *Phaeocystis globosa* bloom was observed along the French coast. The proposed data processing methodology allowed us to define environmental states, to delineate water masses based on (1) the geometry of (unlimited) biotic and abiotic parameters in a new spectral dimension and (2) the spatial and/or temporal dynamics of these states, water masses. When subject to HAB events, these water masses can be easily identified in time and space. Moreover, controlling factors (e.g. nutrient concentrations) and/or direct and indirect effects of these events (e.g. hypoxia) can be hierarchized.

Research in marine ecology (from the description of bio-geochemical equilibria to the prediction of ecosystem dynamics) and implementation of observation and monitoring programs with high temporal and spatial resolution are currently not limited by our potential to develop measurement systems, but by available computing capacities, i.e. by our ability to develop data processing methodologies to manage and process the huge amounts of data generated by these monitoring instruments. The objective for scientists, but also for water managers and stakeholders, should clearly be to integrate high resolution results in their methods of assessing environmental status, improving knowledge of marine ecosystems, answering EU directives (WFD, MSFD) or regional sea convention (OSPAR) issues, and consequently implementing optimized, cost-effective programs and measures to achieve a healthy environment where most of the problems associated with eutrophication (e.g. loss of diversity) do not occur.

LITERATURE CITED

Aiken CM, Petersen W, Schroeder F, Gehrung M (2011) Ship-of-opportunity monitoring of the Chilean fjords using the pocket FerryBox. J Atmos Ocean Technol 28: 1338−1350

Alvarez-Fernandez S, Riegman R (2014) Chlorophyll in North Sea coastal and offshore waters does not reflect long term trends of phytoplankton biomass. J Sea Res 91: 35−44

Aminot A, Kérouel R (2004) Hydrologie des écosystèmes marins: paramètres et analyses. Ifremer, Brest

Anonymous (1988) Courants de marée dans la Manche et sur les côtes françaises de l'Atlantique. Service Hydrographique et Océanographique de la Marine (SHOM), Paris

Beutler M, Wiltshire KH, Meyer B, Moldenke C and others (2002) A fluorometric method for the differentiation of algal populations *in vivo* and *in situ.* Photosynth Res 72: 39−53

Bonato S, Christaki U, Lefebvre A, Lizon F, Thyssen M, Artigas LF (2015) High spatial variability of phytoplankton assessed by flow cytometry, in a dynamic productive coastal area, in spring: the eastern English Channel. Estuar Coast Shelf Sci 154:214−223

Boyce DG, Lewis MR, Worm B (2010) Global phytoplankton decline over the past century. Nature 466:591−596

Boyen C, Heip C, Cury P, Baisnée PF and others (2012) EuroMarine Research Strategy report: deliverable 3.2. Seventh Framework Programme Project EuroMarine − Integration of European Marine Research Networks of Excellence. FP7-ENV-2010.2.2.1−3. Contract Number 265099

Brunet C, Brylinski JM, Bodineau L, Thoumelin G, Bentley D, Hilde D (1996) Phytoplankton dynamics during the spring bloom in the south-eastern English Channel. Estuar Coast Shelf Sci 43:469−483

Brylinski JM, Aelbrecht D (1993) Plankton transfers and coastal front in the Dover Strait. Oceanol Acta 16:671−676

Brylinski JM, Lagadeuc Y (1990) L'interface eaux côtières/

eaux du large dans le Pas-de-Calais (côte française): une zone frontale. CR Seances Acad Sci 311:535–540

Brylinski JM, Lagadeuc Y, Gentilhomme V, Dupont JP and others (1991) Le fleuve côtier: un phénomène hydrologique important en Manche Orientale. Exemple du Pas-de-Calais. Oceanol Acta 11:197–203

Catherine A, Escoffier N, Belhocine A, Nasri AB and others (2012) On the use of the FluoroProbe®, a phytoplankton quantification method based on fluorescence excitation spectra for large-scale surveys of lakes and reservoirs. Water Res 46:1771–1784

Cloern JE (1996) Phytoplankton bloom dynamics in coastal ecosystems: a review with some general lessons from sustained investigation of San Francisco Bay, California. Geophysics 34:127–168

Cloern JE, Jassby AD (2008) Complex seasonal patterns of primary producers at the land-sea interface. Ecol Lett 11: 1294–1303

Crise A, Kaberi H, Ruiz J, Zatsepin A and others (2015) A MSFD complementary approach for the assessment of pressures, knowledge and data gaps in southern European seas: the PERSEUS experience. Mar Pollut Bull 95: 28–39

Cugier P, Billen G, Guillaud JF, Garnier J, Menesguen A (2005) Modelling the eutrophication of the Seine Bight (France) under historical, present and future riverine nutrient loading. J Hydrol (Amst) 304:381–396

Dickey JM (2003) Emerging ocean observations for interdisciplinary data assimilation systems. J Mar Syst 40-41: 5–48

Escoffier N, Bernard C, Hamlaoui S, Groleau A, Catherine A (2015) Quantifying phytoplankton communities using spectral fluorescence: the effects of species composition and physiological state. J Plankton Res 37:233–247

European Parliament and Council (2000) Directive 2000/ 60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Off J Eur Comm L 327:1–72

European Parliament and Council (2008) Directive 2008/ 56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive) (Text with EEA relevance). Off J Eur Union L 164:19–40

Glibert PM (2017) Eutrophication, harmful algae and biodiversity – challenging paradigms in a world of complex nutrient changes. Mar Pollut Bull 124:591–606

Harrison JW, Howell ET, Watson SB, Smith REH (2016) Improved estimates of phytoplankton community composition based on in situ spectral fluorescence: use of ordination and field-derived norm spectra for the bbe FluoroProbe. Can J Fish Aquat Sci 73:1472–1482

Heisler J, Glibert PM, Burkholder JM, Anderson DM and others (2008) Eutrophication and harmful algal blooms: a scientific consensus. Harmful Algae 8:3–13

Hernández-Fariñas T, Soudant D, Barillé L, Belin C, Lefebvre A, Bacher C (2014) Temporal changes in the phytoplankton community along the French coast of the eastern English Channel and the southern Bight of the North Sea. ICES J Mar Sci 71:821–833

Houliez E, Lizon F, Thyssen M, Artigas LF, Schmitt FG (2012) Spectral fluorometric characterization of haptophyte dynamics using the FluoroProbe: an application in the eastern English Channel for monitoring Phaeocystis globosa. J Plankton Res 34:136–151

Lancelot C (1995) The mucilage phenomenon in the continental coastal waters of the North Sea. Sci Total Environ 165:83–102

Lefebvre A, Guiselin N, Barbet F, Artigas LF (2011) Long-term hydrological and phytoplankton monitoring (1992–2007) of three potentially eutrophic systems in the eastern English Channel and the Southern Bight of the North Sea. ICES J Mar Sci 68:2029–2043

Legendre L, Legendre P (1998) Numerical ecology. Elsevier, Amsterdam

MacIntyre HL, Lawrenz E, Richardson TL (2010) Taxonomic discrimination of phytoplankton by spectral fluorescence. In: Suggett DJ, Prášil O, Borowitzka MA (eds) Chlorophyll a fluorescence in aquatic sciences: methods and applications. Springer, Dordrecht, p 129–169

Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (2002) The Pacific Decadal Oscillation. J Oceanogr 58: 35–44

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? J Classif 31:274–295

Muylaert K, Gonzales R, Franck M, Lionard M and others (2006) Spatial variation in phytoplankton dynamics in the Belgian coastal zone of the North Sea studied by microscopy, HPLC-CHEMTAX and underway fluorescence recordings. J Sea Res 55:253–265

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: Dietterich TG, Becker S, Ghahramani Z (eds) NIPS '01: Proc 14th Int Conf Neural Info Process Syst: Natural and Synthetic, 3–8 December 2001,Vancouver. MIT Press, Cambridge, MA, p 849–856

OSPAR Commission (2009) Eutrophication status of the OSPAR maritime area: second OSPAR integrated report. Eutrophication Series, Publication No. 372/2008. OSPAR Commission, London

OSPAR Commission (2010) Quality status report 2010. OSPAR Commission, London

Petersen W, Schroeder F, Engelke C (2006) A mature system for operational monitoring. Sea Technol 47:53–57

Petersen W, Wehde H, Kraseman H, Colijn F, Schroeder F (2008) FerryBox and MERIS – assessment of coastal and shelf sea ecosystems by combining in situ and remotely sensed data. Estuar Coast Shelf Sci 77:296–307

Phan TTH, Poisson-Caillault E, Lefebvre A, Bigand A (2017) Dynamic time warping-based imputation for univariate time series data. Pattern Recognit Lett, doi:10.1016/ j.patrec.2017.08.019

R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rantajärvi E (ed) (2003) Alg@line in 2003: 10 years of innovative plankton monitoring and research and operational information service in the Baltic Sea. MERI – Report Series of the Finnish Institute of Marine Research No. 48. Finnish Institute of Marine Research, Helsinki

Richardson TL, Lawrenz E, Pinckney JL, Guajardo RC, Walker EA, Paerl HW, MacIntyre HL (2010) Spectral fluorometric characterization of phytoplankton community composition using the Algae Online Analyser® Water Res 44:2461–2472

Rousseau V, Mathot S, Lancelot C (1990) Calculating carbon biomass of Phaeocystis sp. from microscopic observations. Mar Biol 107:305–314

Rousseau V, Lancelot C, Cox D (eds) (2006) Current status of

eutrophication in the Belgian coastal zone. Presse Universitaires de Bruxelles, Bruxelles

Rousseeuw K, Poisson-Caillault E, Lefebvre A, Hamad D (2015) Hybrid hidden Markov model for marine environment monitoring. IEEE J Sel Top Appl Earth Obs Remote Sens 8:204–213

Ruser A, Popp P, Kolbowski J, Reckermann M and others (1999) Comparison of chlorophyll-fluorescence based measuring systems for the detection of algal groups and the determination of chlorophyll-*a* concentrations. Berichte Forsch Technologiezentr Westküste Univ Kiel 19: 27–38

Salomon JC, Breton M (1993) An atlas of long-term currents in the Channel. Oceanol Acta 16:439–448

Sazhin AF, Artigas LF, Nejstgaard JC, Frischer ME (2007) The colonization of two *Phaeocystis* species (Prymnesiophyceae) by pinnate diatoms and others protests: a significant contribution to colony biomass. Biogeochemistry 83:137–145

Schroeder F, Mizerkowski B, Petersen W (2008) The pocket-FerryBox—A new portable device for water quality monitoring in oceans and rivers. J Oper Oceanogr 1: 51–57

See JH, Campbell L, Richardson TL, Pinckney JL, Shen R (2005) Combining new technologies for determination of phytoplankton in the northern Gulf of Mexico. J Phycol 41:305–310

Seppälä J, Olli K (2008) Multivariate analysis of phytoplankton spectral *in vivo* fluorescence: estimations of phytoplankton biomass during a mesocosm study in the Baltic Sea. Mar Ecol Prog Ser 370:69–85

Seuront L (2005) Hydrodynamic and tidal controls of small-scale phytoplankton patchiness. Mar Ecol Prog Ser 302: 93–101

Shumway RH, Stoffer DS (2006) Time series analysis and its applications (with R examples), 2nd edn. Springer, New York, NY

Smayda TJ (1990) Novel and nuisance phytoplankton blooms in the sea: evidence for a global epidemic. In:

Granéli E, Sundström B, Edler L, Anderson DM (eds) Toxic marine phytoplankton. Elsevier, New York, NY, p 29–41

SRN (Suivi Régional des Nutriment) (2017) SRN dataset - regional observation and monitoring program for phytoplankton and hydrology in the eastern English Channel. 1992-2016. SEANOE http://doi.org/10.17882/50832

Teixeira H, Berg T, Uusitalo L, Fürhaupter K and others (2016) A catalogue of marine biodiversity indicators. Front Mar Sci 3:207

Thyssen M, Alvain S, Lefebvre A, Dessailly D and others (2015) High-resolution analysis of a North Sea phytoplankton community structure based on in situ flow cytometry observations and potential implication for remote sensing. Biogeosciences 12:4051–4066

Utermöhl H (1958) Zur vervollkommnung der quantitativen phytoplankton-methodik. Mitt Int Ver Theor Angew Limnol 9:1–38

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Yentsch CS, Yentsch CM (1979) Fluorescence spectral signatures: the characterization of phytoplankton populations by the use of excitation and emission spectra. J Mar Res 37:471–483

Zelnik-Manor L, Perona PP (2004) Self-tuning spectral clustering. In: Saul LK, Weiss Y, Bottou L (eds) NIPS '04: Proc 17th Int Conf Neural Info Process Syst, 13–18 December 2004, Vancouver. MIT Press, Cambridge, MA, p 1601–1608

Zongo S, Schmitt F (2011) Scaling properties of pH fluctuations in coastal waters of the English Channel: pH as a turbulent active scalar. Nonlinear Process Geophys 18: 829–839

Zongo S, Schmitt F, Lefebvre A, Repecaud M (2011) Observations biogéochimiques des eaux côtières à Boulogne-sur-Mer à haute fréquence: les measures automatiques de la bouée MAREL. In: Schmitt FG (ed) Observation des écosystèmes marins et terrestres de la Côte d'opale: du naturalisme à l'écologie. Presses de l'UOF, Paris, p 253–266